Capstone Project- 1

Airbnb Booking Analysis

Team Name:

Vijay

Team member's name:

Pushpendra Saini

Ayush Singh

Debaprasad Mohapatra

Contents:

Problem Statements

- **Exploratory Data Analysis on the Dataset that consists details of bookings on Airbnb Platform.**

- **Objectives of the EDA:**

  - **To know the variables and performing some analysis on those variables to get some insights.**
  - **Data Visualization**
  - **Data cleaning, outlier treatment and feature engineering to prepare the data for the creation of machine learning models.**
  - **Looking at the correlations**
  - **Finding the conclusions**

Inspiration

- What can we learn about different hosts and areas?
- What can we learn from predictions? (ex: locations, prices, reviews, etc)
- Which hosts are the busiest and why?
- Is there any noticeable difference of traffic among different areas?

Data Summary

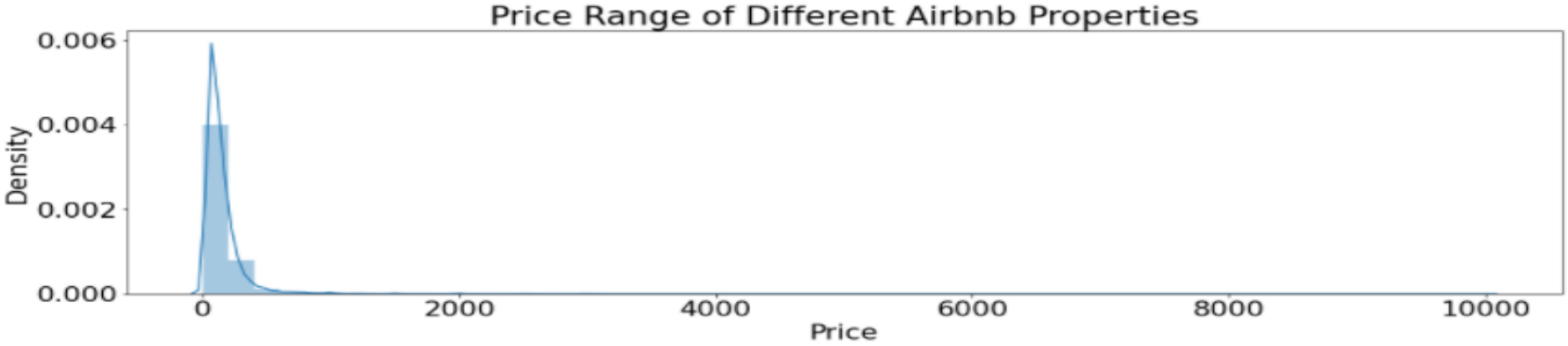
- **Data set name:** Airbnb Bookings Analysis
- **Shape:**
  - Rows:48895
  - Columns:16
- **Important Variables:**

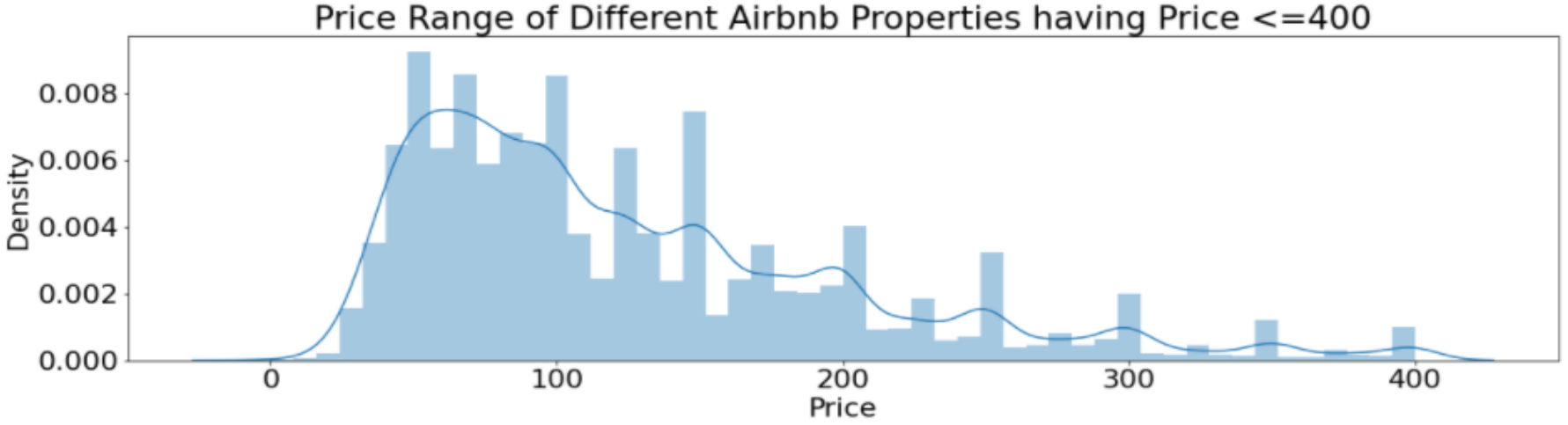
- id: This variable contains unique id for a unique listing.
- name: This gives us a small introduction about a listing. For example: Cozy Clean Guest Room - Family Apt, Large Furnished Room Near B'way etc.
- host_id: This variable contains id for a host who is the host of corresponding listing.
- host_name: This gives us name of the host of corresponding listing in the same row.

- neighbourhood_group: New York City is composed of five boroughs: The Bronx, Brooklyn, Manhattan, Queens, and Staten Island. These boroughs are named here as neighbourhood_group.
- neighbourhood: These are the areas in their corresponding neighbourhood_group.
  room_type: This variable tells us about the type of listings. There are three types of Listings, 'Private room', 'Entire home/apt', 'Shared room'.
- price: This is the price in dollars for one night stay.
- minimum_nights: minimum nights, someone can book that listing for.
- number_of_reviews: Total number of reviews for that listing.
- last_review: This shows date of the latest review.
- reviews_per_month: number of reviews per month for that listing.
- calculated_host_listings_count: number of listings listed per host
- availability_365: number of days when listing is available for booking out of 365 days.

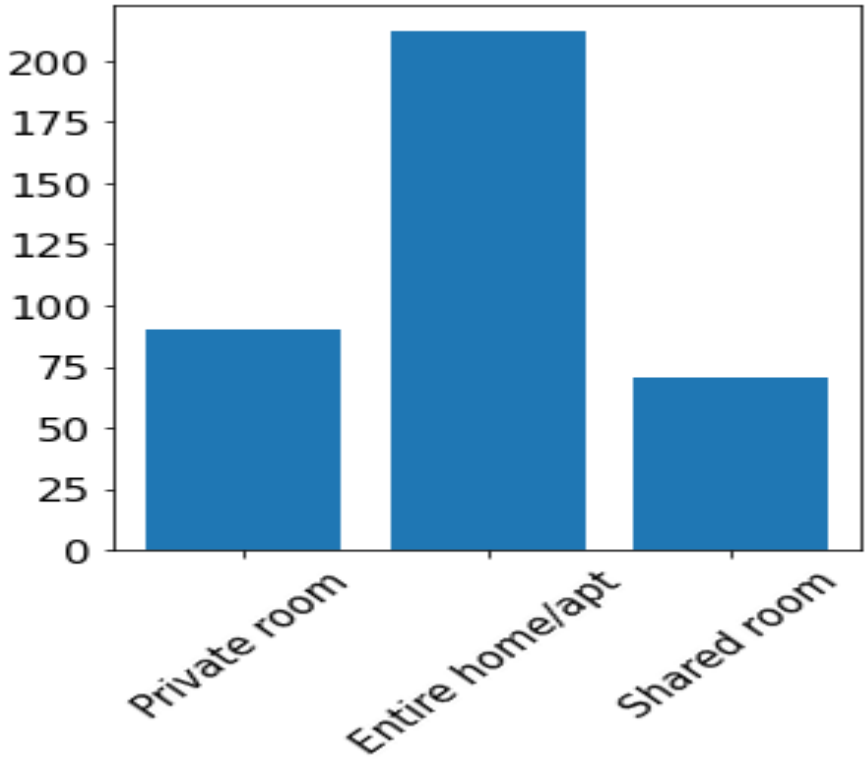
- We changed name of some columns:
  - 'id'- 'listing_id', 'name'- 'listing_details', 'minimum_nights'- 'minimum_nights_stay', 'last_review'- 'last_review_date', 'availability_365'- 'booking_availability'

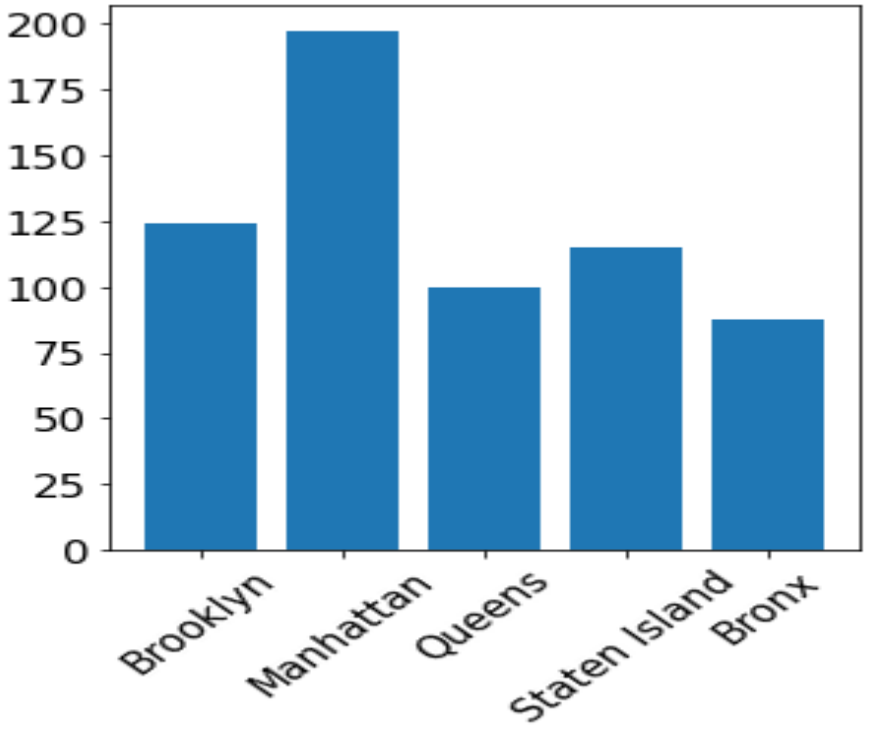
Analysis on price column



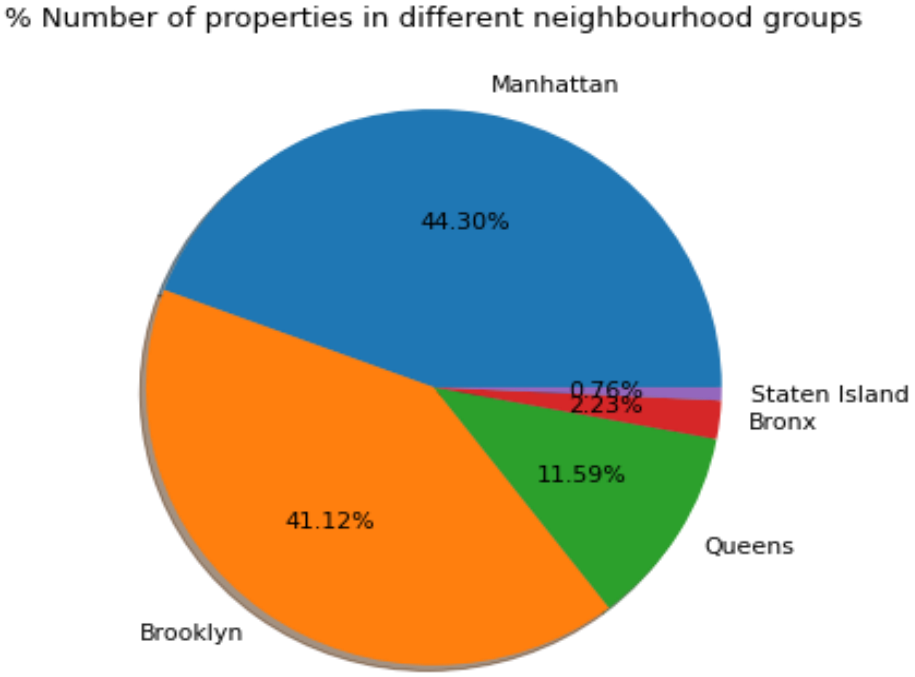


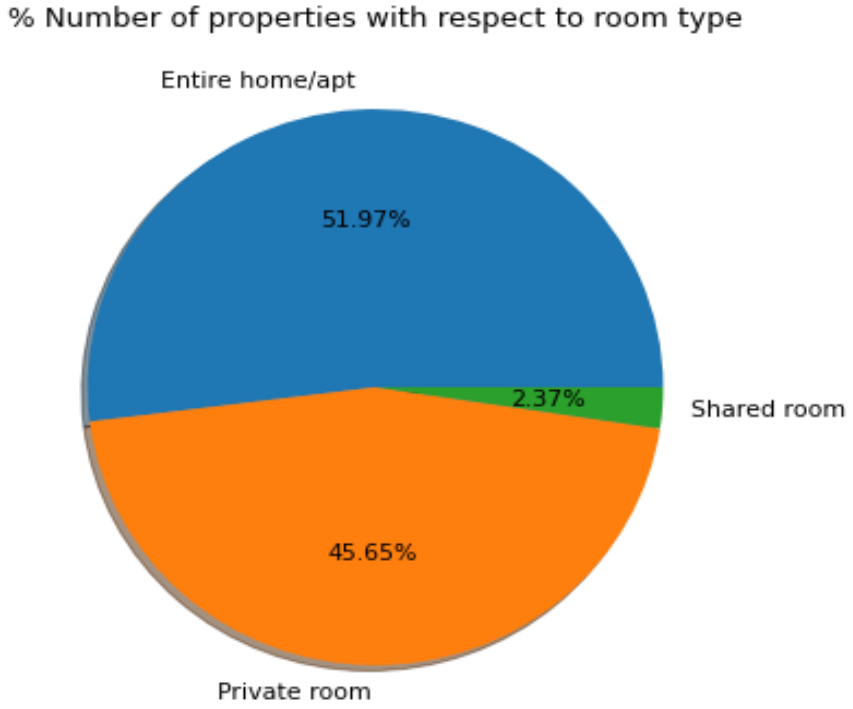
Average price according to room type and neighbourhood group



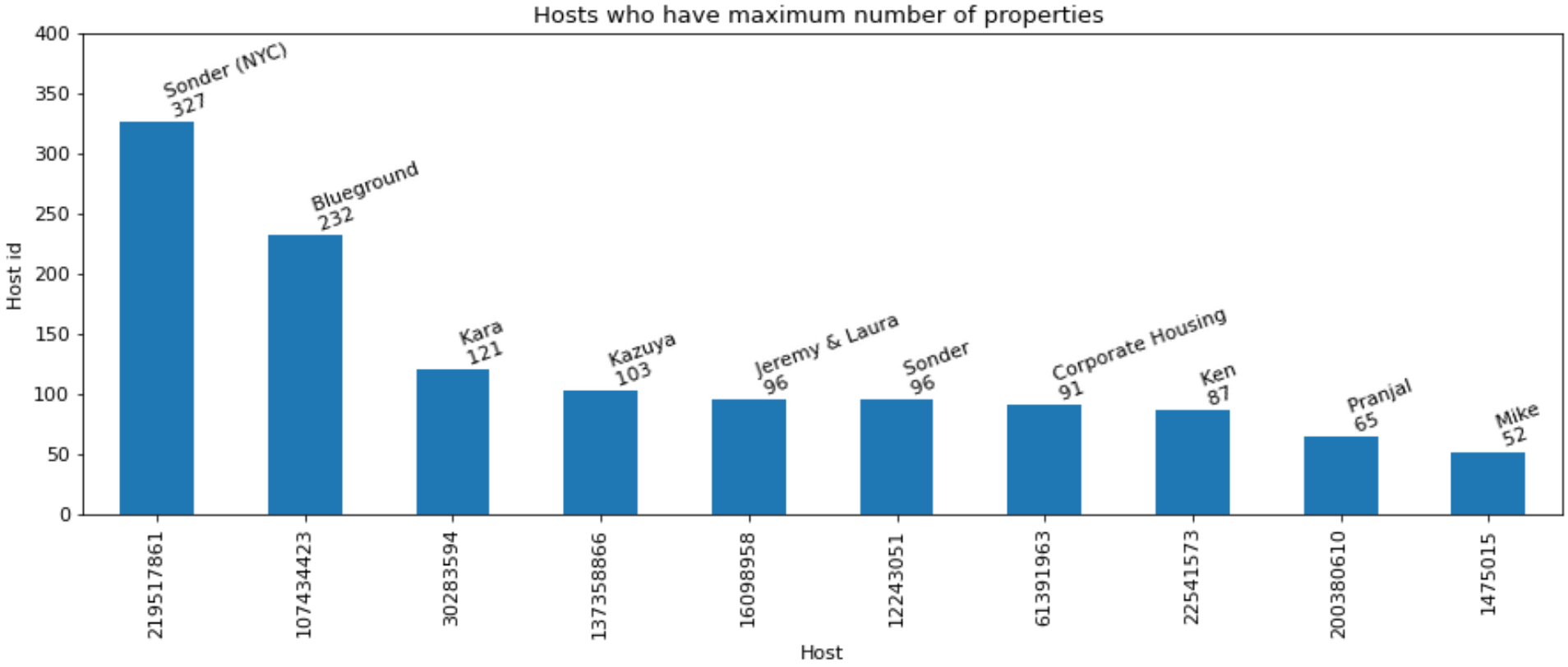


Neighbourhood groups and room types

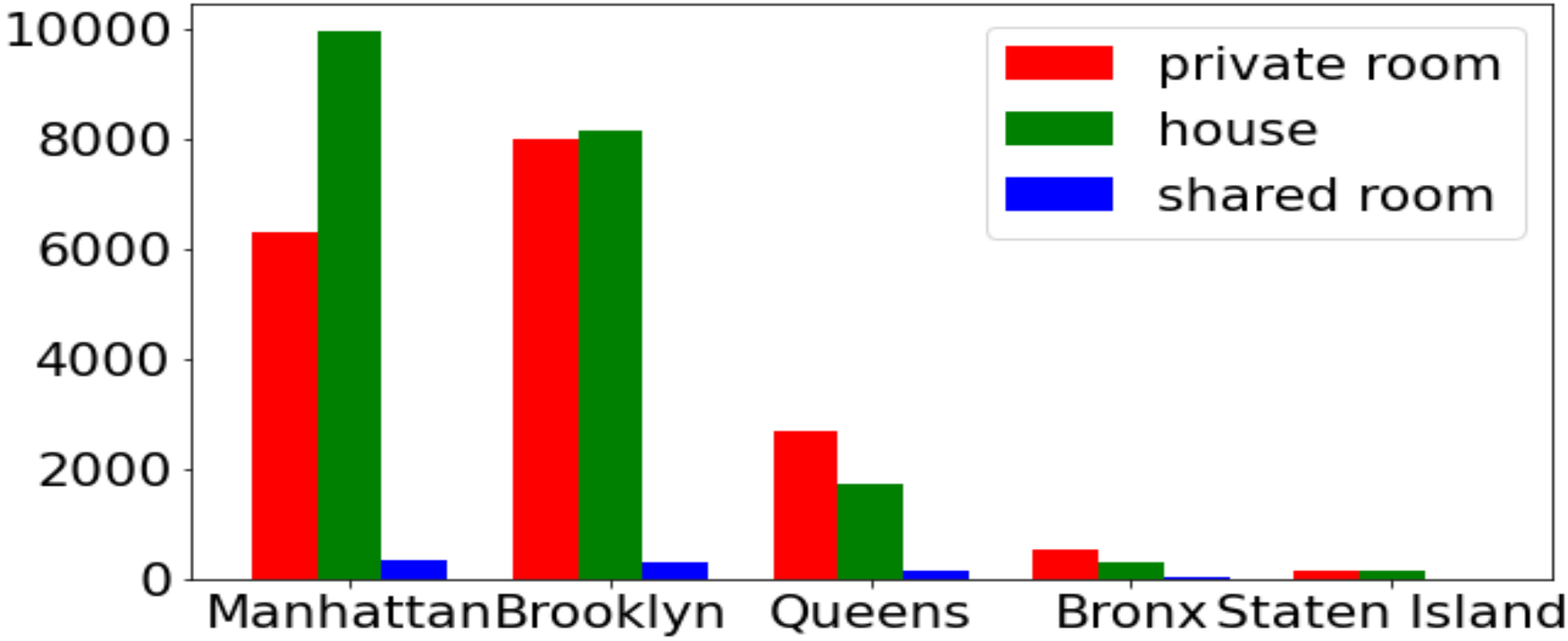




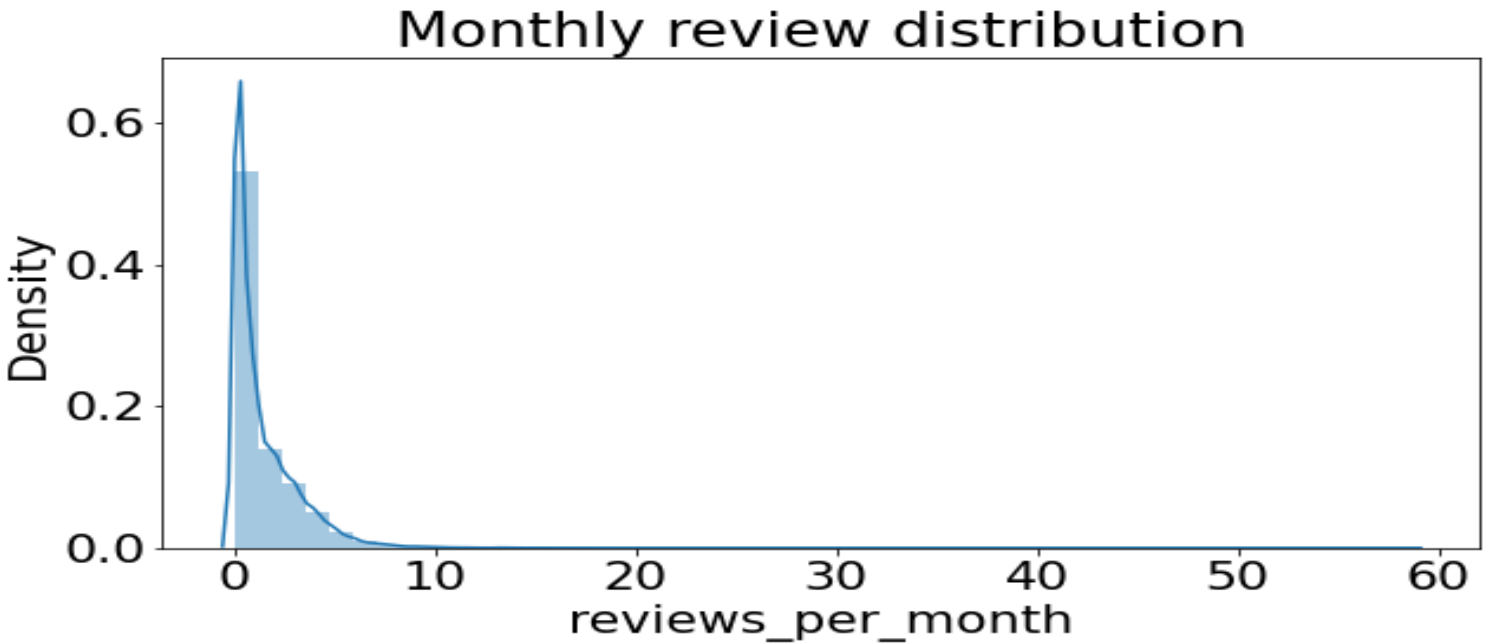
Hosts that hold the maximum number of properties



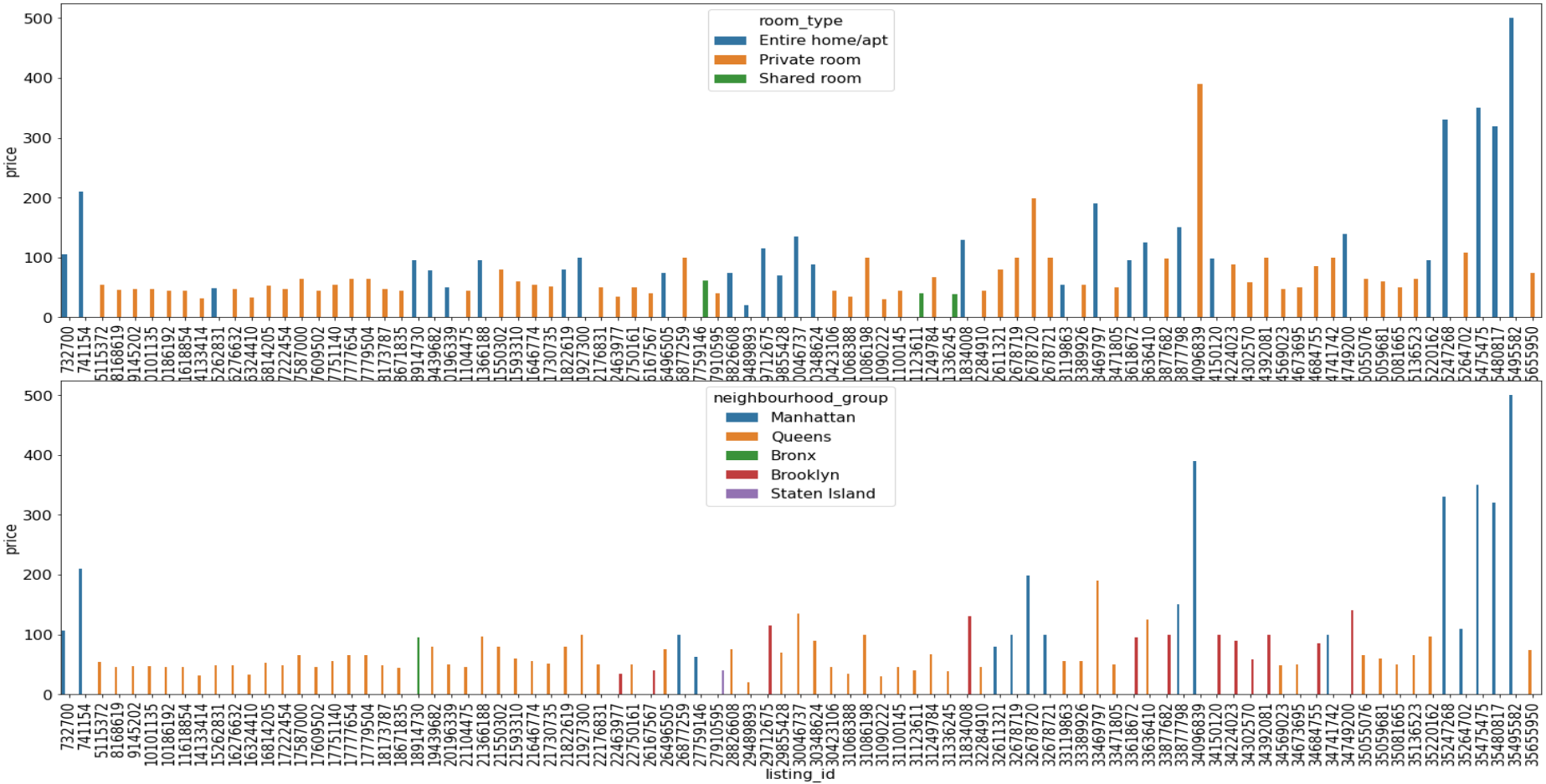
Types of room in different neighbourhod groups



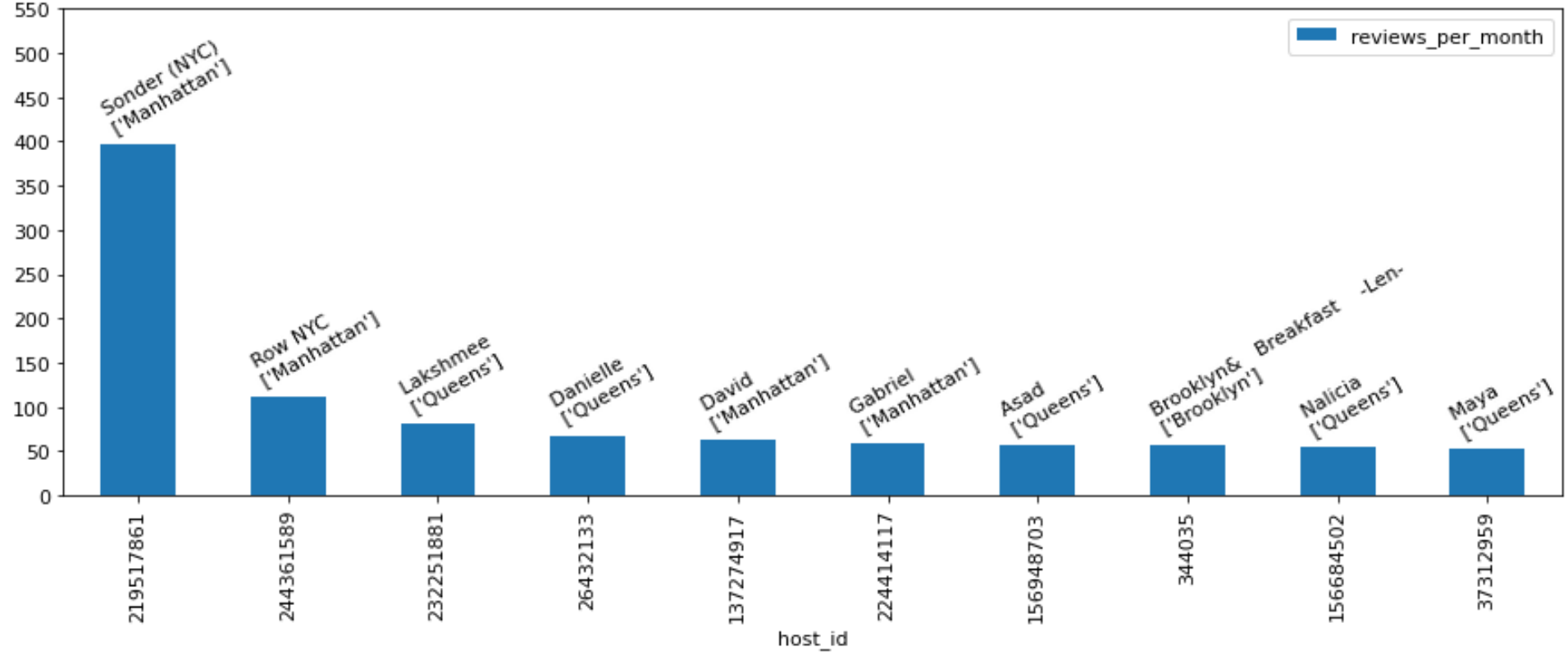
Analysis on monthly review



Properties with good number of reviews



Top 10 busiest hosts

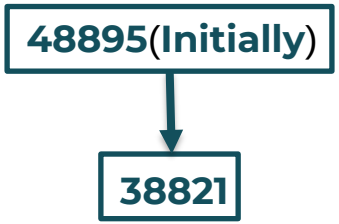


Data Cleaning

- **Look at the info of Dataset**

0	listing_id	48895 non-null	int64
1	listing_details	48879 non-null	object
2	host_id	48895 non-null	int64
3	host_name	48874 non-null	object
4	neighbourhood_group	48895 non-null	object
5	neighbourhood	48895 non-null	object
6	latitude	48895 non-null	float64
7	longitude	48895 non-null	float64
8	room_type	48895 non-null	object
9	price	48895 non-null	int64
10	minimum_nights_stay	48895 non-null	int64
11	number_of_reviews	48895 non-null	int64
12	last_review_date	38843 non-null	object
13	reviews_per_month	38843 non-null	float64
14	calculated_host_listings_count	48895 non-null	int64
15	booking_availability	48895 non-null	int64

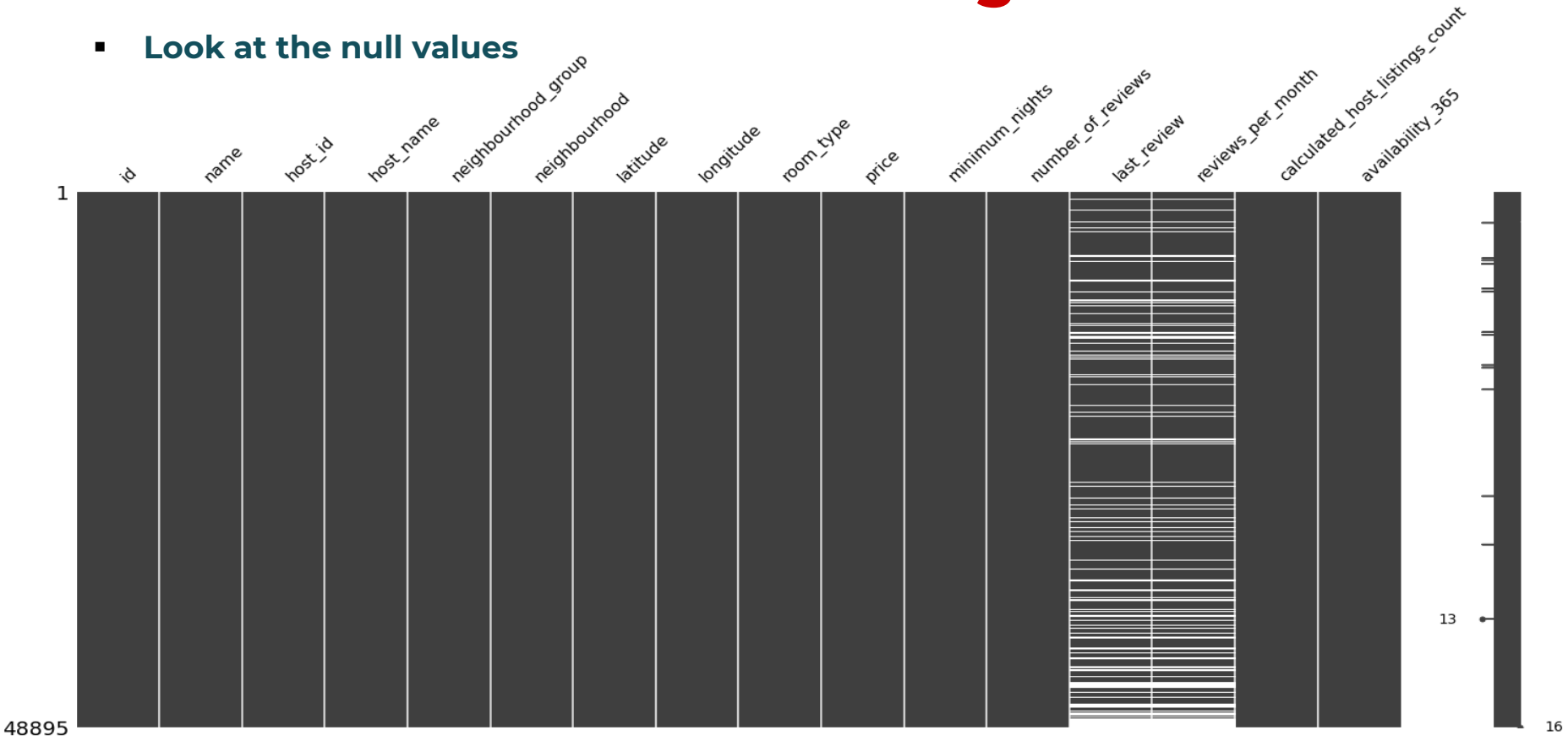
- **Removing rows with at least one null value**



- **Huge loss of data, so worked with original Dataset**

Data Cleaning

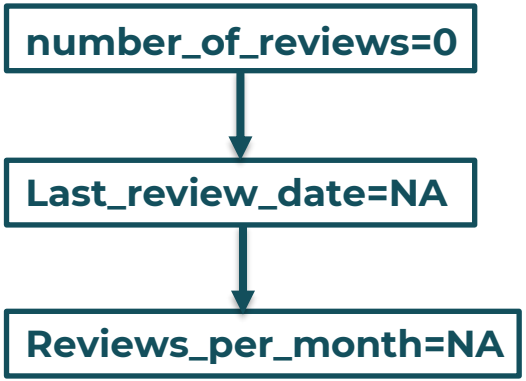
- **Look at the null values**



Data Cleaning

- **Look at the NA values in last_review and reviews_per_month column :**

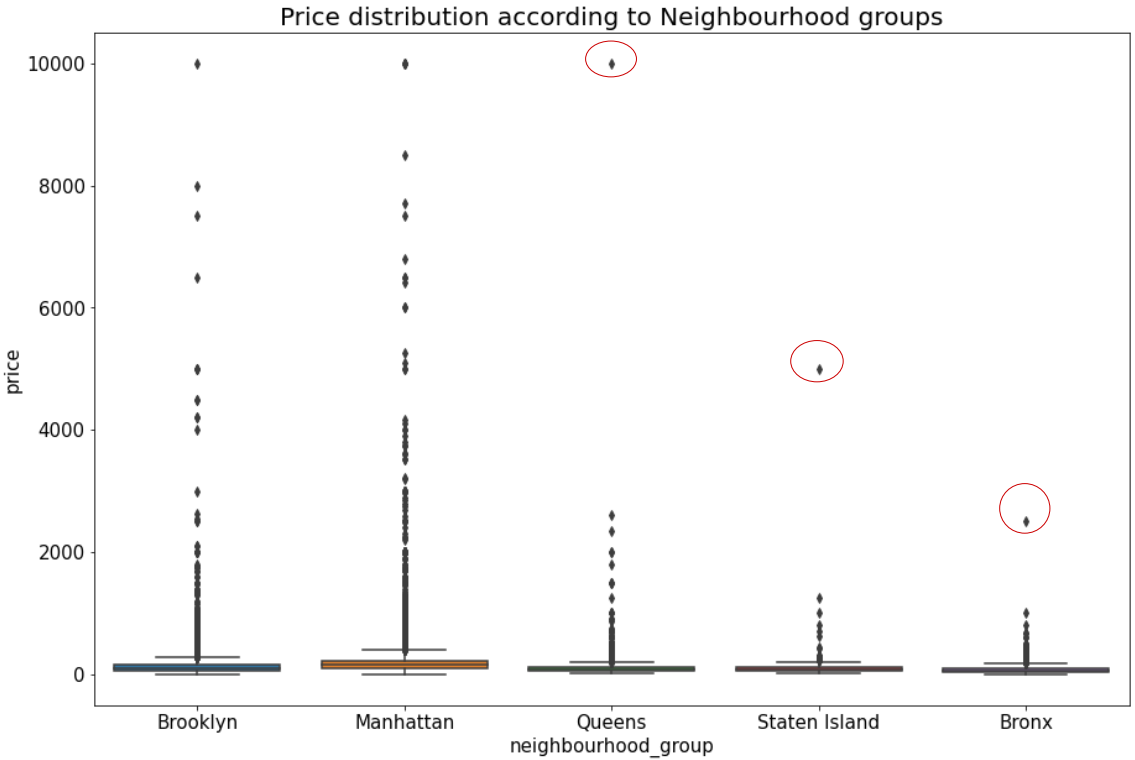
number_of_reviews	last_review_date	reviews_per_month
0	NaN	NaN
0	NaN	NaN
0	NaN	NaN
0	NaN	NaN
0	NaN	NaN



- **No changes in these columns**
- **Changed value zero with median value in price column**

Outlier Detection & Treatment

- **Look at the box plot for price distribution for each neighborhood group :**



- **Removed outliers having unusual price in comparison with other listings in the same neighborhood group (In the red circles)**

Outlier Detection & Treatment

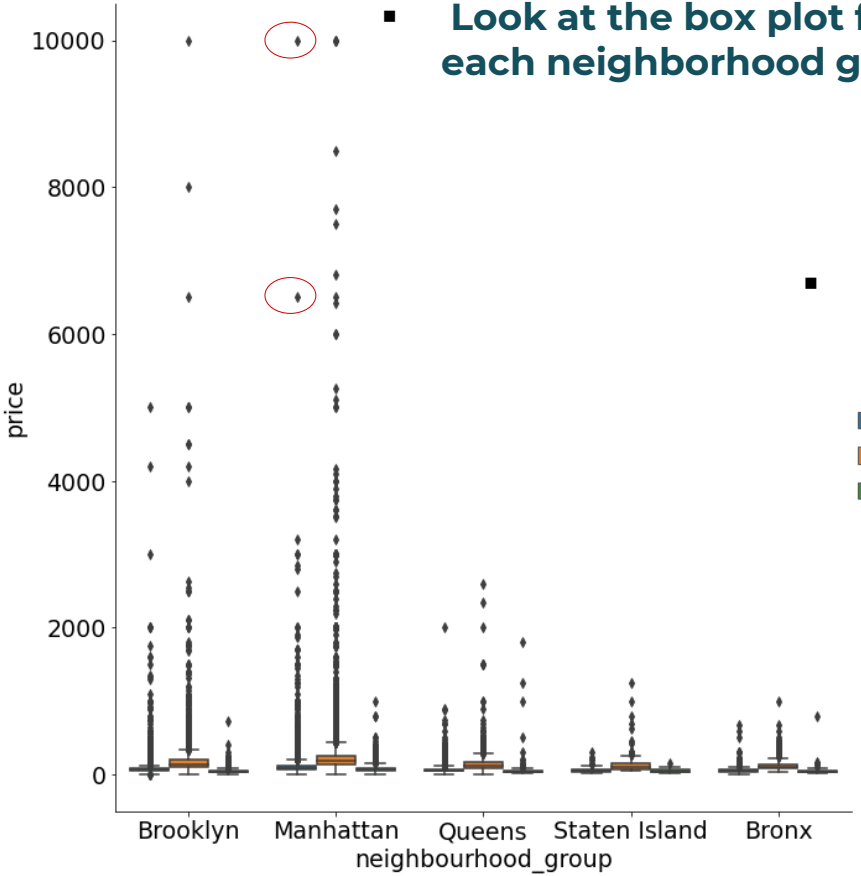
- look at the listings for the hosts who have at least one listing in >7000 $ Range.

	listing_id	listing_details	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights_stay
45554	34830282	Clarkson Loft the gem of east Flatbush	262534951	Sandra	Brooklyn	East Flatbush	40.65904	-73.92334	Private room	60.0	1
45666	34895693	Gem of east Flatbush	262534951	Sandra	Brooklyn	East Flatbush	40.65724	-73.92450	Private room	7500.0	1

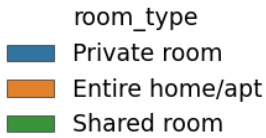
- Removed second row with 7500 $ price.

Outlier Detection & Treatment

- Look at the box plot for price distribution for different type of rooms in each neighborhood group :



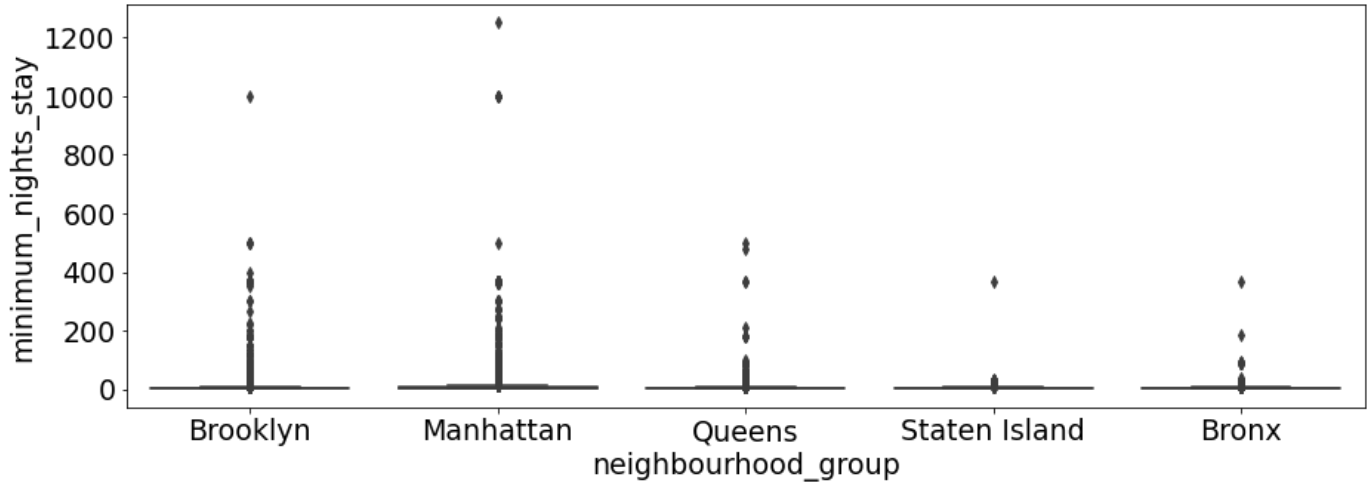
- Changed room type of two listings from private room to entire home/apt.



Outlier Detection & Treatment

- Look at the box plot for minimum nights stay in each neighborhood group
- Look at the table. Angie, one of the host, have one property with 999 nights, different from others. Replaced this with 30.
- Look at the values of minimum nights stay that are more than 365 nights, replaced all values with 365.

370	1000	1250	500	500	480	500	999	500	366	999	400	999	500



Host	Minimum_nights_stay
Angie	999
Angie	30
Angie	30
Angie	30
Angie	30
Angie	30

Creating features

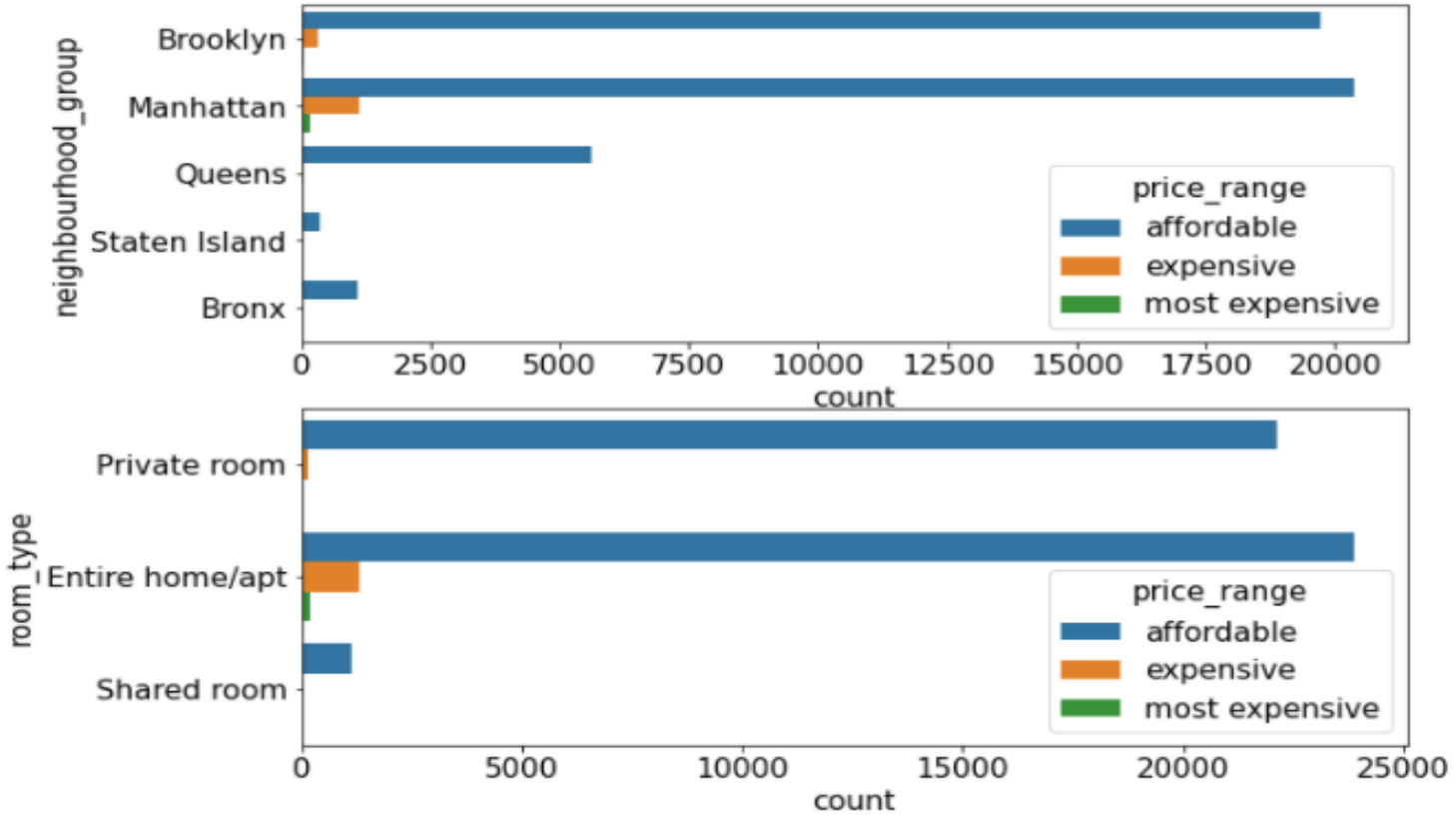
price_range	Entire home/apt	Private room	Shared room	availbility_cat
affordable	0	1	0	high availability
affordable	1	0	0	high availability
affordable	0	1	0	high availability
affordable	1	0	0	average availability
affordable	1	0	0	not available

We created a new feature for price, giving some labels according to price range. Most listings are in price range of 0 to 400 dollars.
so we gave a label to this range: affordable
400 to 1000 dollars: expensive
price >1000 dollars: most expensive

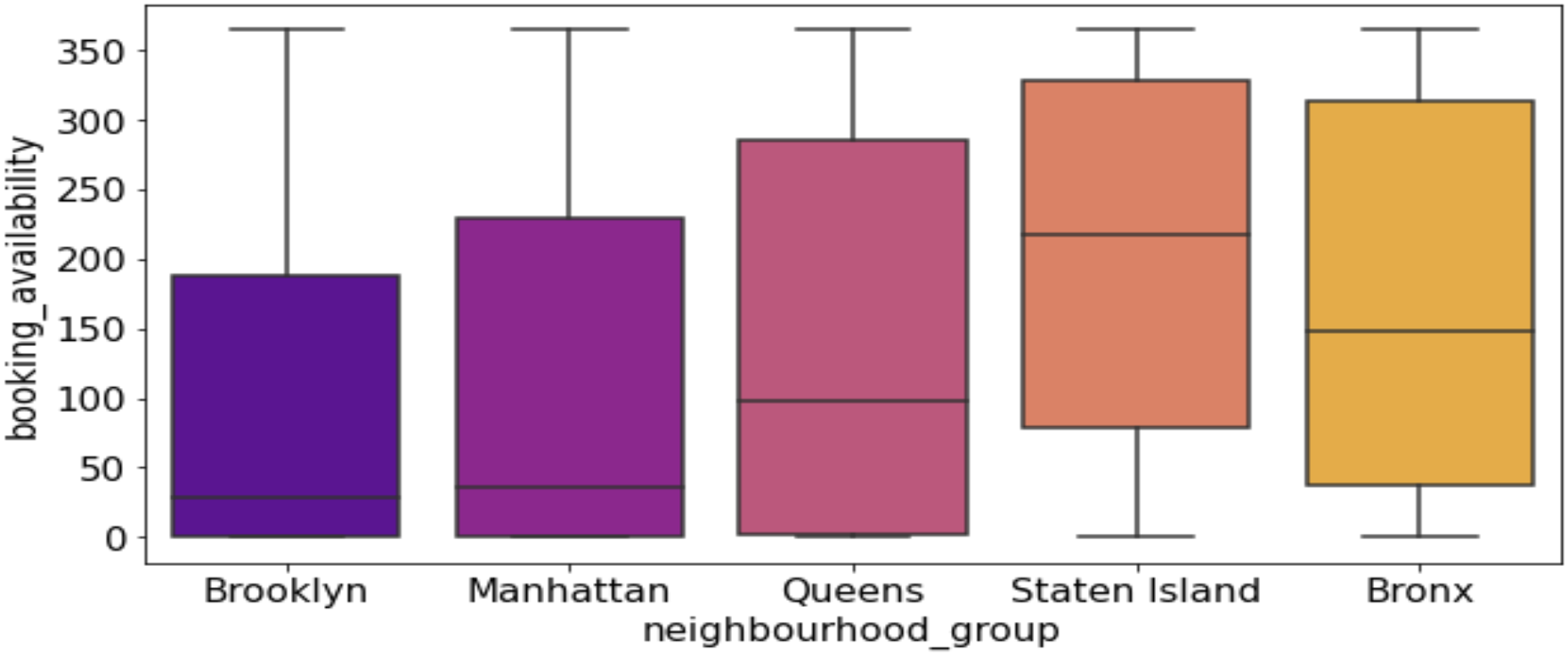
We created dummy variables for room type.
This will help us in using this feature while creating Machine Learning models. And the numeric values help to see if there is any correlation.

Now to create a new feature for booking availability we have to divide availability in proper categories. In our data density is more for availability<100. and same is with availability >300. So we created four bins:
availability 0 days: not available
availability 0<days<= 90: low availability
availability 90<days<=270: average availability
availability days>270: high availability

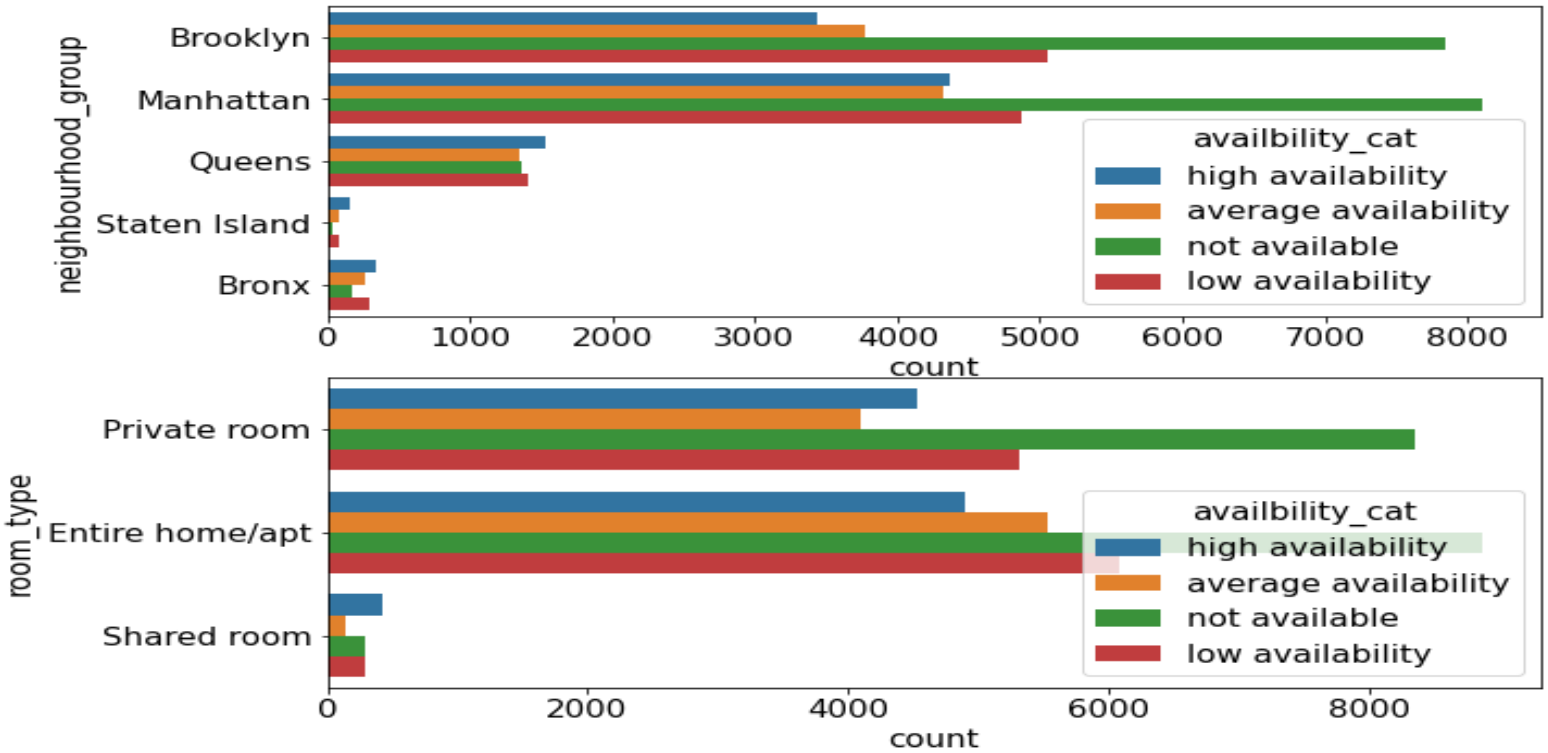
Categorical feature for price



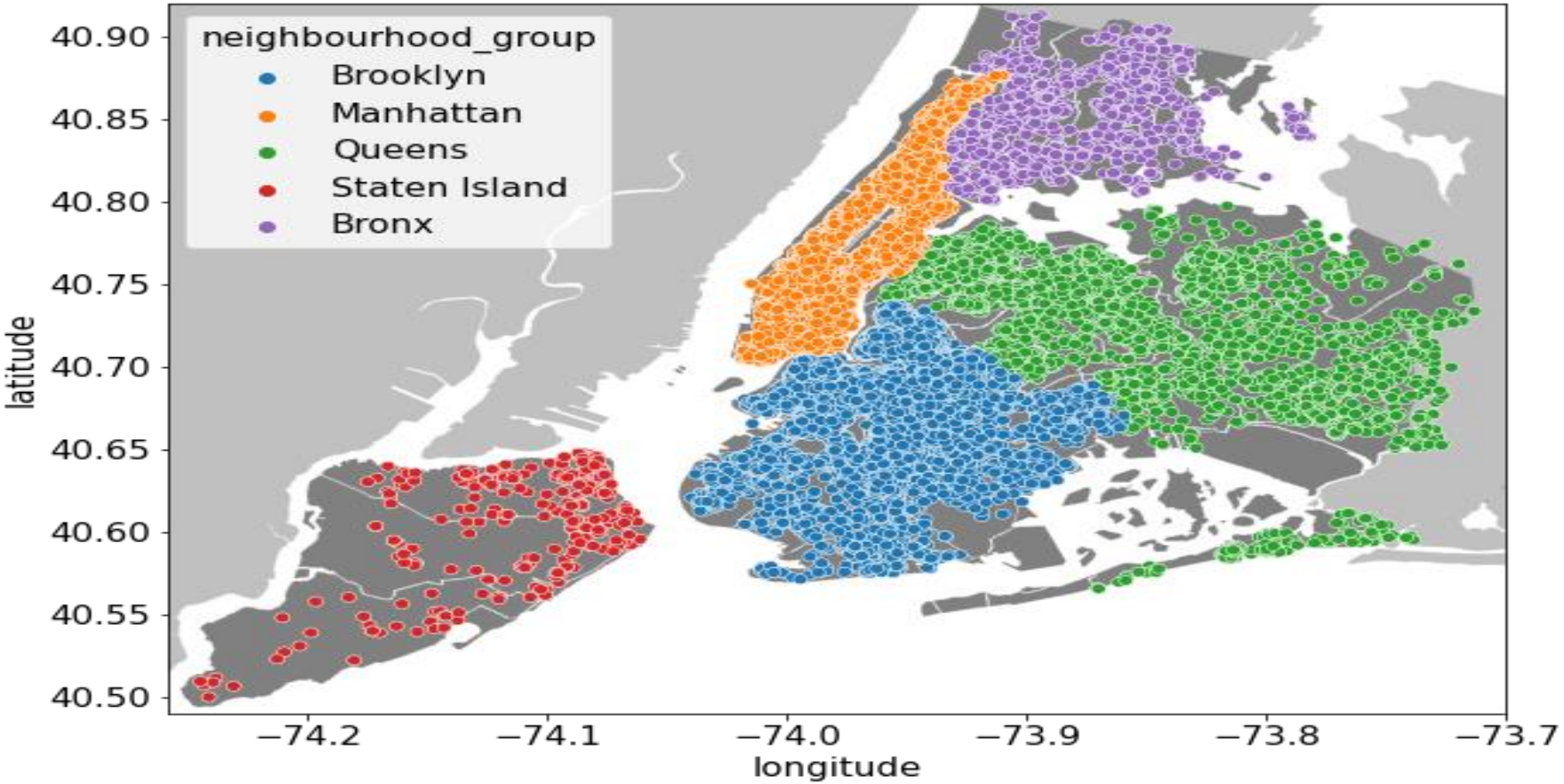
Booking availability in neighbourhood group



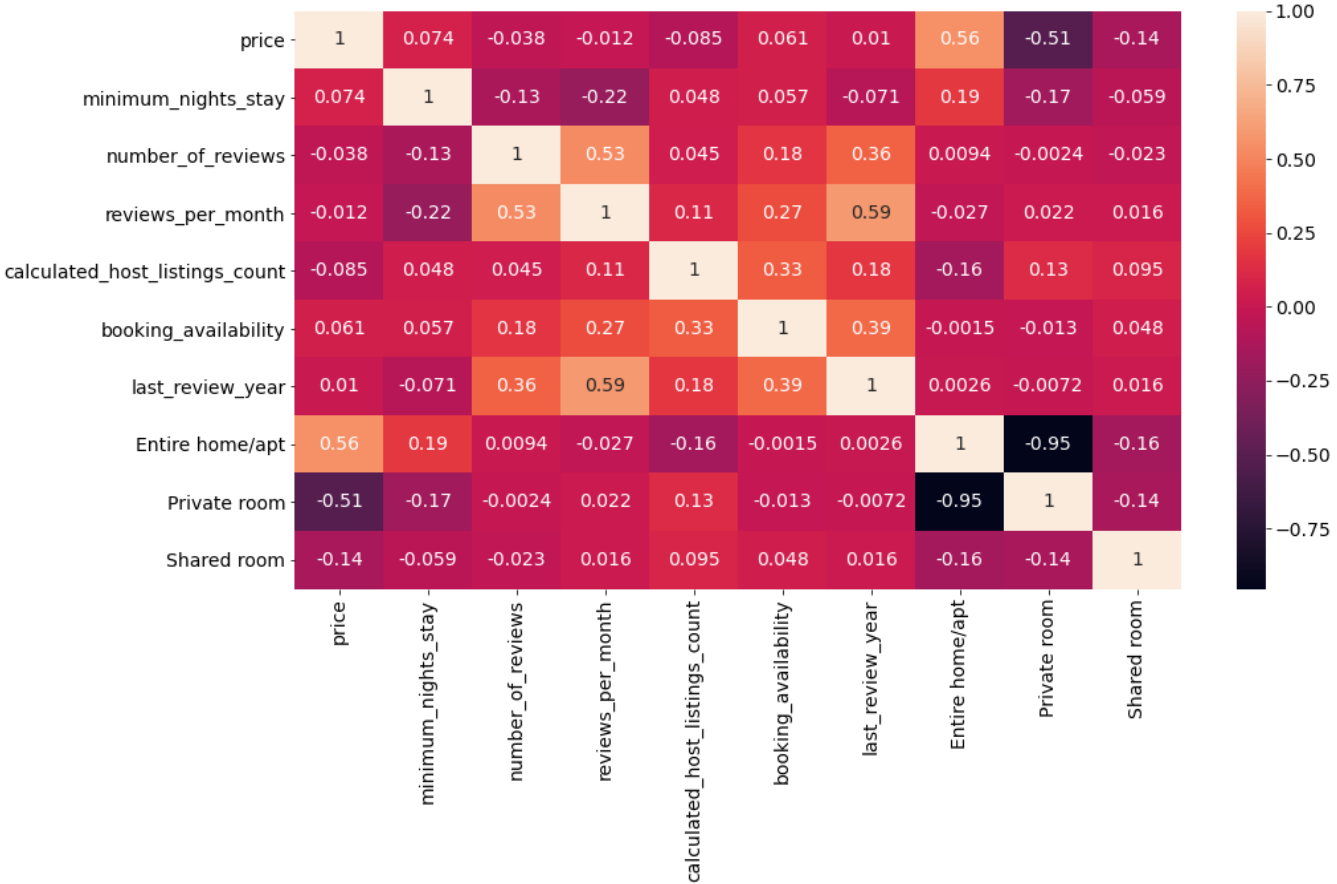
Availability of room types in neighbourhood group



Map of neighbourhood groups



Heatmap


price 1 0.074 -0.038 -0.012 -0.085 0.061 0.01 0.56 -0.51 -0.14
minimum_nights_stay 0.074 1 -0.13 -0.22 0.048 0.057 -0.071 0.19 -0.17 -0.059
number_of_reviews -0.038 -0.13 1 0.53 0.045 0.18 0.36 0.0094 -0.0024 -0.023
reviews_per_month -0.012 -0.22 0.53 1 0.11 0.27 0.59 -0.027 0.022 0.016
calculated_host_listings_count -0.085 0.048 0.045 0.11 1 0.33 0.18 -0.16 0.13 0.095
booking_availability 0.061 0.057 0.18 0.27 0.33 1 0.39 -0.0015 -0.013 0.048
last_review_year 0.01 -0.071 0.36 0.59 0.18 0.39 1 0.0026 -0.0072 0.016
Entire home/apt 0.56 0.19 0.0094 -0.027 -0.16 -0.0015 0.0026 1 -0.95 -0.16
Private room -0.51 -0.17 -0.0024 0.022 0.13 -0.013 -0.0072 -0.95 1 -0.14
Shared room -0.14 -0.059 -0.023 0.016 0.095 0.048 0.016 -0.16 -0.14 1
price
minimum_nights_stay
number_of_reviews
reviews_per_month
calculated_host_listings_count
booking_availability
last_review_year
Entire home/apt
Private room
Shared room
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Challenges

- Less correlation between the variables.
- We might need more features to build machine learning models.
- Lack of domain knowledge.
- Detecting the outliers.

Conclusion

- Out of five boroughs Manhattan and Brooklyn are the most expensive boroughs.
- Most of the properties listed (around 96%) have been priced under 400 dollars.
- Sonder(NYC), Row(NYC), Lakshmee, Danielle and David are the top 5 busiest hosts. In the top ten hosts, most of the hosts have their properties in Manhattan and Queens.
- There is no strong correlation between numeric variables. To create a better machi ne learning model to predict price we might need other information's. For example: per capita income, locations of airports, stations, number of rooms, type of furnishing etc.
- Some properties are having good reviews so Airbnb can take feedbacks from the corresponding hosts about their extra initiatives and share to others to attract more customers.
- There are three type of rooms present, in which Entire home/apt has the highest frequency followed by private rooms then shared rooms. Entire home/apt are the most expensive and the shared rooms are the cheapest ones.

Contributor role

Name: Pushpendra Saini
Email i'd: sainipushpendra08@gmail.com
Contribution: Data Cleaning and Outlier treatment

Name: Ayush Singh
Email i'd: asa494013@gmail.com
Contribution: Univariate, Bivariate and Multivariate analysis

Name: Debaprasad Mohapatra
Email i'd: debaprasad13@gmail.com
Contribution: Creating New Features, Data Visualization and Finding Correlation

Q & A